大規模高可用性サーバの設計と運用

敷田 幹文　井口 寧　三輪 信介
丹 康雄　松澤 照男

北陸先端科学技術大学院大学 情報科学センター

近年，インターネットの普及に伴い，電子化される業務範囲が拡大したため，データの量や重要度が急速に増大し，各業務の情報ネットワーク依存度が高くなってきている．各サーバメーカでは，サービスの可用性を高めるための各種のハードウェアおよびソフトウェアを備えているが，大規模データベースなどの特殊用途を考えたものであり，多くのユーザが日常的に利用するファイル共有や電子メールサービスのサーバに利用するには課題も多い．本論文では，著者らの大学においてこれまでに導入した大規模高可用性サーバシステムの設計方針について述べ，また，それぞれのシステムのこれまでの運用経験をもとに，高可用性サーバの設計法と運用法の指針を提案する．

The Design Method of Large-scale High Availability Servers

Mikifumi SHIKIDA　Yasushi INOGUCHI　Shinsuke MIWA
Yasuo TAN　Teruo MATSUZAWA

Japan Advanced Institute of Science and Technology

Recently, because transaction area in local area network expanded caused by the spread of the Internet, the amount and importance of data increase rapidly, and the information network dependency of each transaction has been rising. However there are many problems to use past high availability hardware and software for general services such as file sharing and e-mail. In this paper, we describe the design of high availability servers in our campus network. We propose the guideline of the design method of high availability server systems based on our actual operational servers.

1　はじめに

近年，インターネットの普及に伴い，電子化される業務範囲が拡大し，データの量や重要度が急速に増大している．そのため，組織内ネットワークは大規模化，集中化する傾向にあり，また，各業務の情報ネットワーク依存度が高くなってきている．その結果，オープンシステムを用いる一般の企業や大学でも，多数のクライアントにサービスを提供するサーバとして，これまで以上に信頼性の高いシステムが求められている．

これに対するソリューションとして，各サーバメーカでは，可用性を高めるための各種のハードウェアおよびソフトウェアを備えている．しかし，これらの商品はこれまで大規模データベース等特殊な領域で用いられてきたため，ファイル共有サーバや電子メールサーバ等の一般的なサービスのサーバとしての歴史はあまりなく，大規模なサーバとして構築した公開されている事例も少ない．そのため，これらの個々の商品を組み合わせた全体システムとしての構成法が確立されておらず，設計・運用のためには特殊な知識と経験が要求されていた．

本論文では，著者らの大学においてこれまでに導入した大規模高可用性サーバシステムの設計について述べ，また，それぞれのシステムのこれまでの運用経験をもとに，高可用性サーバの設計法と運用法の指針を提案する．

以下，２章では既存の高可用性システムの特徴

について述べ，3章で我々の大学における高可用性NFSファイル共有サーバの設計法を述べ，4章でそれらのシステムの運用事例を通して高可用性システムについての議論を行う．

2 高可用性システム

本章では，これまで商品化されている主な高可用性システムが，どのような仕組みを用いているかを整理し，既存の高可用性システムの特徴について述べる．

2.1 高可用性システムとは?

システムが提供するサービスは，各種の障害やメンテナンスなどによって中断する可能性があるが，代替機や代替部品によってその中断時間を通常よりも短くすることが可能である．このように設計されたシステムを，本論文では「高可用性(High Availability)システム」と呼ぶ．

可用性を高めるためには，ハードウェアの各部およびソフトウェアが二重化や冗長化のための機能を持っているが，その範囲は広く，どの部分が必須であるといった一般的な基準は特にない．

- ディスク装置内部

  大型サーバは大容量の二次記憶装置を必要とすることが多く，大量のディスクドライブを持つ．ディスクドライブは故障の頻度が高いため，複数のドライブを組み合わせて冗長な書き込みを行い故障時にもデータの破壊を防ぐRAIDやホットスワップなどの機構が従来から用いられてきた．通常これらはディスクアレイコントローラによって制御される．

  しかし，近年，RAID機能を持ったディスクアレイ装置は比較的一般的なものとなってきているため，本論文ではこの部分の冗長化機能を持っているだけでは高可用性システムとは呼ばない．

- ディスク装置とCPU本体の接続

  ディスク装置内のディスクアレイコントローラはRAIDの制御やキャッシュなど複雑な機能を持っており，このボードが故障することも少なくない．複数のコントローラボードを用いて，故障時の自動代替機能を実現している装置も多い．また，CPU本体側のSCSIインタフェースボードと複数のコントローラを繋ぐSCSIバスも複数本に独立させることが可能なシステムもあるが，障害時にはアクセスするバスが替わることになるため，CPU側でもこれを制御するソフトウェアが必要となる．

- CPU本体

  CPU本体を二重化しておくことにより，CPU本体のハードウェアやその上で動くOSやアプリケーションに障害があった場合にも他のCPUが代替(takeover)することが可能となる．その場合，アプリケーションの環境を代替機が引き継ぐ必要がある．多くのシステムでは，TCP/IPを用いてクライアントに何らかのサービスを行っているため，IPアドレスの引継ぎ機能が必要となる．その際，引継ぎ用に独立したネットワークインタフェースボードを持つシステムと，同一のインタフェースボードに複数のIPアドレスをエイリアスとして付け，代替機が通常行っているサービスと代替したサービスを両立させるシステムがある．なお，本論文では，このCPU本体の代替機能を持っていることが高可用性システムの重要な用件と考える．

- CPU本体とネットワークの接続

  CPU本体全体の代替機能とは別に，1台の本体の中でネットワークインタフェースボードのみの故障の際に，ボード単位に代替を可能とするシステムも多い．その場合，ボードからルータまでの経路も二重化できるため，その間のケーブル断線などの障害にも対応できるが，ルータ側でもVLANなどを用いてこの二重化に対応する必要がある．

- サーバ・クライアント間のネットワーク

  実際にサーバの運用を行うためには，サービスがクライアント機から利用できなければ意味をなさない．そのため，サーバとクライアントを繋ぐ組織内ネットワークの各所でも可用性を高める機構が必要になる．

2.2 高可用性システムの特徴

高可用性システムとして構築したサーバは，その機能を持たない従来のサーバと比較して以下のような特徴をもつ．

障害時対処の時間的余裕 ほとんどのシステムでは，障害発生時に自動的に障害箇所を検出し，稼

動可能部分が代替する機能を有している．従って，夜間や休日など管理者が不在の状況で障害が発生した場合でも対処を翌日以降に遅らせることができる．

24時間サービスを行うシステムでは，管理を担当する者の個人的時間にまで影響をおよぼすことがあるため，管理者の精神的負担に大きな違いがある．特に，大学では，システムを大型機から移行し，サービスに対する要求の増大によって24時間サービスを行うことが増えたが，人件費などの制度面の改善が遅れているためこの利点の意味は非常に大きい．

メンテナンスによる停止の減少 OSのパッチ当てなど，システムの計画的メンテナンスの場合にも，CPU本体の代替機能を用いることによって，切り替え時の2-3分以外はサービスを継続したまま多くのメンテナンスが可能である．これによって，例えば平日昼間のサービス中に行うことが可能になるため，管理者側やメーカ側技術者の人員確保など態勢が整った状態でのメンテナンス作業が可能となる．

システム価格の上昇 高可用性システムとして構成するには，構成可能な装置に制限があることが多い．共有二次記憶装置やネットワーク部分の二重化箇所を増やし，また，サービス中に交換可能な箇所を増やすに従って価格が上昇する．一般に，2台のクラスタを構成する際には，1台のサーバの価格の2倍よりかなり高価になると言える．

システムの複雑化 高可用性システムでは，周辺機器を含めたハードウェア及びソフトウェアが特殊な機能を持っており，通常のシステムに較べて複雑なものとなる．それらは各メーカや製品によって異なる部分があり，システム導入後の管理者のシステム習得や管理ノウハウの蓄積にも労力を要する．また，導入事例が少ないために，ハードウェアやソフトウェアの不具合も多い．これはソフトウェアの実現上の不具合も多いが，仕様に問題点があることも少なくない．さらに，障害が発生した場合に，障害に対処できる業者側の技術者が少ない．

3 高可用性サーバの設計

前章で述べた各部の設計は，対象となるサーバの目的やサービスの特徴付けによって必要な個所や設定が異なる．本章では，著者らの大学における，NFSファイル共有サーバ数種の事例を交えて，高可用性サーバの設計法について述べる．

3.1 情報環境 FRONTIER

"FRONTIER"は北陸先端科学技術大学院大学の情報ネットワーク環境の総称であり，具体的には，個人用計算機，ファイルサーバ群，計算サーバ群，その他の各種サーバ群，および，これらを接続するネットワーク等からなる．

FRONTIERは最初の学生が入学した1992年4月に合わせて稼働を開始し，現在では1,300人以上のユーザと7,000以上のホスト[1]が接続されているが，1つのドメインで集中管理し，情報科学センターが全体の設計・運用・管理を行っている．これによって，TCOを削減し，また計算機を専門としない一般ユーザにも同一環境を提供している[7]．

各ユーザのホームディレクトリやアプリケーション等はほとんどをNFSファイルサーバに格納しているが，それらのサーバのディスクは，1992年の開学時から全てRAID5またはRAID1で構成し，運用してきた．1996年以降，順次，高可用性システムに更新している．

一方，ネットワークは，学内を2つの領域に分割し，バックボーンルータは2台で，それぞれにメインとサブを用意しているため，計4台の大型の高速レイヤ3スイッチが接続されている．

3.2 設計事例

本学でこれまでに導入した高可用性NFSサーバの一覧を表1に示す．これ以外にも，メールサーバ，WWWサーバなどに高可用性システムを導入している[2]．これらのシステムの中で，代表的構成のシステムについて以下に述べる．

3.2.1 基本的な構成例

ファイルサーバFS5aは本学で最初に導入した高可用性システムである．挑戦的システムではあったが，初めての試みであるため，代替可能箇所の割合が小さく，問題点も少なくなかった．

筐体は1本のラックにCPUとディスク等が収まっており，CPU本体はレールに乗っていてメンテナンス時には引き出すことによってメンテナンス

[1]DHCPを除くNISエントリ数

表 1: 本学で導入した高可用性 NFS ファイルサーバの一覧

略称	導入年	本体	ディスク	ソフトウェア
FS5a	1996	SONY	SONY	FirstWatch[6]
FS3b	1998	Sun	Sun	SunCluster[5]
FS0c	1999	Sun	Sun	SunCluster
FS7a	1999	IBM	IBM	HACMP[3], HANFS[4]
FS1c	2000	富士通, PFU, NetworkAppliance		ONTAP[8] (NFS 専用機)
FS5b	2000	HP	日立	MC/ServiceGuard[1]

性を向上させる筐体になっている．しかし，そのためにシステム稼動状態でケーブル等が接続されたままボード交換ができず，高可用性システムとしては逆にメンテナンス性が悪化したことになる．

実際，バックアップ装置との接続箇所が故障した際に，混雑期であったために交換のためのシステム停止が可能になるまで期間があり，同種のバックアップソフトウェアを稼動させている別のサーバでバックアップを行うという方法を用いて運用で回避した．

また，1 つの SCSI バスに 2 つの CPU が接続されるので，マルチイニシエータ接続に対応した SCSI ボードを利用したが，システム導入時点ではその場合の書き込みキャッシュに対応しておらず，クライアントからの書き込み速度が 1/3 以下に落ちた．同一 SCSI バス上のファイルシステムは必ず一方の CPU からしかアクセスしないような運用を行うことで回避した．CPU が 1 台の場合には書き込みキャッシュが働くのが一般的であるので見落としがちであるが，2 台の場合のキャッシュは技術的に難しく，機能しないボードは他のメーカでも存在する．

一方，ディスクや CPU 本体などに障害が発生した場合の通知は，管理コンソール上に表示される．目立つ GUI 表示やアラーム音によって工夫されているが，この種の大型サーバは，筐体の大きさや騒音の点からも管理者の居室とは別室に設置されることが多いため，このような機能だけでは管理者に通知されない．また遠隔監視ツールがあっても，システムごとに別になると多数のシステムの管理は難しくなる．我々は，障害発生時に電子メールによって通知する機構を実現した．

3.2.2 レイヤ 2 スイッチを用いた構成例

ここでは，ファイルサーバ FS3b を例として，レイヤ 2 スイッチを介してルータと接続した構成について述べる．

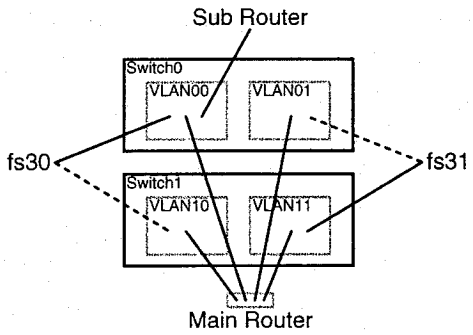

図 1: ファイルサーバ FS3b のネットワーク接続

前に述べたように FRONTIER のネットワークには主・副の 2 つのバックボーンルータがあり，サーバはこの両方に接続する必要がある．FS3b の CPU が持つネットワークインタフェースはトランキングや VLAN を組むような方法に対応していない．1 つのセグメントに複数インタフェースを用いる機構もあるが，同時に使われるのは 1 個で他はスタンバイなので，稼働している 1 つのインタフェースから 2 つのルータに接続する必要がある．

この実現のためにレイヤ 2 スイッチを用いて分岐させる方法が一般的であるが，単純な接続ではスイッチの故障時にルータとの接続性が失われてしまう．即ち，サーバもルータも共に高可用性設計がされているにもかかわらず，スイッチが可用性を下げてしまう．

この問題を解決するため，FS3b では図 1 のような接続とした．なお，この図は 1 つのセグメントのみを示しており，実際にはこのような接続が複数ある．これにより，スイッチの一方が完全に停止しても，本体側はインタフェースの冗長機構によって利用するインタフェースボードが切り替わり，ルータ側も VLAN により同一アドレスになっており，MAC アドレスの再学習が行なわれ，平常時と同様の通信が行なわれる．なお，ルータ間のループを防ぐために，バックアップルータへの接

続は4つのうち1つのみとなっている．そのためバックアップルータへの通信には平常時にもメインルータを通過することになる．メインルータが完全に停止した場合には，4つのVLANのうち1つのみがバックアップルータに接続された状態となるが，インタフェースボードの切り替えとCPU本体のtakeoverが発生し，この1つのみでサービスが継続可能である．

また，ここで用いたスイッチは2台をスタック可能で，複数ポートのsharing等も可能であるため，この他の構成として，4つのVLANを1つにすることによってバックアップルータへの接続性を確保し，メインルータへの帯域幅も増やす方法など，いくつかの構成が考えられたが，各部での障害時の動作を検討した結果，スイッチ・ルータ側の動作が最も単純な本構成を採用することとした．一般に，サーバ上の特にネットワークに関係した部分で複雑な機構を導入している場合に，ネットワーク機器など別のレイヤでも複雑な構成にすると，障害発生時の解析・対処が困難になると言える．特に，販売数が少ないシステムでは，他メーカの機器との接続事例も少なく，障害時にメーカ側でも想定していなかった状態となって不具合が露見することも多い．

障害の通知は，本システムにおいても管理コンソール上でGUIを利用したツール上で表示されるため，外部に対して能動的に通知を行うことはない．そのため，syslog機能と連携させて，管理者・業者へメールによる障害通知を行う機構を実現した．

一方，重要なユーザデータを扱うファイルサーバでは，テープメディアなどへのバックアップ機構が重要であるが，クラスタ制御機構と連係して片肺運転時にもバックアップを行なうように構成することは極めて難しく，連係機能を持ったシステムは少ない．本システムで採用したSunClusterとバックアップソフトウェアのNetBackupの場合でも，別企業で開発されたソフトウェアであり連係機能はない．しかし，我々は，SunClusterとNetBackup両方で相手の存在を考慮するように設定し，片肺運転時にもバックアップすることを可能とした．これに関する詳細は別の機会に述べる．

3.2.3 レイヤ3スイッチを用いた構成例

ここでは，ファイルサーバFS5bを例として，レイヤ3のスイッチングルータを利用してネットワーク接続した構成について述べる．

FS5bでは，前述の4つのルータへ接続するために，図2に示すように二重化した1組のルータを用いている．ただし，図2では4つのルータとの接続は省略している．VLANはセグメントに対応して用意する．2台のルータはトランキングによる高帯域のネットワークで繋がれており，各VLANはtagを付けてルータ間を通すことにより，全セグメントを両方のルータに用意している．

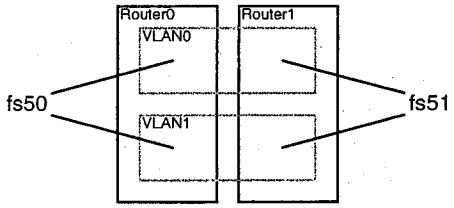

図2: ファイルサーバFS5bのネットワーク接続

このような接続によって，一方のルータがダウンした場合にも他方のルータにより全セグメントがサービスされる．CPUからは1つのセグメントに1本のケーブルしか出ないため，ダウンしたルータに繋がっていたインタフェースは利用できなくなるが，ルータとCPUを1対1に対応させることにより，ルータダウン時の影響を片方のCPUに集中させているため，サーバ側も自動的にtakeoverしてサービスが継続される．即ち，FS5bはFS3bと違いネットワークインタフェースボード単体での代替機能は持たせていないが，このように構成することによって，ボードやネットワーク側の故障時にもサービスが継続できる．

もし代替ネットワークインタフェースボードが利用可能であれば，FS3bの構成と同様に，fs50とRouter1というように通常状態のボードの接続先と異なる方のルータへ接続しておくことにより，一方のルータが停止した際にCPU本体の代替ではなくボード単位の代替だけで済むため，より早い切り替えが期待される．

4 ディスカッション

高可用性システムの導入前は，障害時に即対処することが求められ，夜間や休日などに対処することも多かった．また，ハードウェア故障時に業者を待つ余裕がないため，保守部品を学内で保管したり，ある程度の部品交換作業はセンター教官・技官が対応することも多かった．ハードウェア故障以外のシステムダウンの場合でも，障害発生時から対処完了後までサービスが停止する．本学の場合，全学のほとんどのワークステーションがアプリケーショ

ンやホームディレクトリのアクセスにNFSファイル共有を行っているため，全学の教育研究活動に致命的な影響を与える．実際，1台のファイルサーバが停止しても電話などによる問い合わせが多くあり，対処に時間を要する場合には掲示を出すなどの手段を必要としていた．他の大学や企業などでも，今後ますます大規模集中化が進むと考えられ，このような運用では教育・研究・営利活動に大きな影響を及ぼすと言える．

これに対して，高可用性システムへの移行後は，管理者不在の状況においても自動的に代替できた障害がほとんどである．CPU本体の切り替えは，速いシステムで1分程度，長いシステムでも2,3分程度で完了する．CPUのダウンによってファイルシステムにダメージがあった場合にはfsckのためにさらに長い時間を要するが，それ以外の通常の切り替え時にユーザから苦情の連絡が来たことはほとんどない．

全ての障害の修理がサービス中に行えるわけではないが，修理のためのサービス停止をユーザ数の少ない日時に設定し，予めユーザにアナウンスできることは，ユーザ側から見たサービスの可用性を考えると大きな違いがあると言える．

一方で，システムが複雑化したため，発生する障害も複雑なものとなってきている．同様のシステムを導入している組織が世界規模で数えても少ないために，発生した障害が開発元でまだ確認されておらず，これに対する対応も次のソフトウェアリリースを待たなければいけないことが度々ある．

また，これらの高可用性システムは，システム単体としてはある程度の完成度が実現されているが，他社製品であるネットワーク機器やクライアントとの接続まで含めた統合システムとして考慮されてはいない．そのため，ネットワーク機器との接続を含めたシステム全体の構成は導入する大学側主導で行い，サーバ側のマニュアルや事例にはない特殊な構成もほとんどは著者らが提案・設計したものである．

5 おわりに

本論文では，著者らの大学における大規模高可用性サーバの設計と運用経験をもとに，このようなサーバシステムの設計法と運用法の指針を提案した．組織内のサーバは，今後ますます大規模化・集中化し，サービスの稼働率の向上など高信頼性が要求される．本論文で述べた方法により，従来の高可用性サーバ用ハードウェア・ソフトウェアと高性能ルータ機を組み合わせ，電子メールやファイル共有などの現在広く一般に利用されているサービスのサーバとして大規模高可用性サーバを構築することが容易になるであろう．

しかし，現在の高可用性サーバは，1つのシステム内で障害を検出してサービスの代替を行っているに過ぎない．そのため，障害発生の通知などをシステム外へ行うには個別の作りこみを必要としている．今後は，全システムを統一的に扱える障害通知プロトコルを確立し，障害情報の収集・分析を行う機構が必要と考えている．

謝辞

本研究を進めるに当たって，各サーバの構築や日常の管理業務では本学情報科学センター技官の方々にご協力頂いております．ここに深く感謝致します．

参考文献

[1] HEWLETT-PACKARD COMPANY. *Managing MC/ServiceGuard*, 1998.

[2] Y. Inoguchi, M. Shikida, Y. Tan, and T. Matsuzawa. Designing networks for large scale distributed systems. In *DIMACS Workshop on Robust Communication Networks*, Nov. 1998.

[3] International Business Machines Corporation. *High-Availability Cluster Multi-Processing for AIX Administration Guide Version4.3.1*, 1999.

[4] International Business Machines Corporation. *High-Availability Cluster Multi-Processing for AIX HANFS for AIX Version4.3.*, 1999.

[5] SUN Microsystems, Inc. *SUN Cluster2.2 System Administration Guide*, 1999.

[6] VERITAS Software Corporation. Veritas FirstWatch インストレーション/環境設定ガイド, 1999.

[7] 敷田幹文, 井口寧, 丹康雄, 松澤照男. 大規模分散システムの集中運用管理における効率化技術の提案. 情報処理学会 分散システム/インターネット運用技術シンポジウム, pp. 75–80, Feb. 1999.

[8] 富士通株式会社/株式会社PFU. NR1000シリーズ ネットワーク接続型ディスク・アレイ装置システム管理者ガイド, 1999.